Pediatric Pain, Predictive Inference, and Sensitivity Analysis

By

Robert Weiss*

School of Statistics, University of Minnesota
Technical Report #596
October, 1993

*Robert Weiss is an Assistant Professor in the Department of Biostatistics at the UCLA School of Public Health, Los Angeles, CA 90024-1772, electronic mail : rob@oahu.ph.ucla.edu. This work was partially supported by NIH grants #MH 37188 and #GM50011. Part of this work was completed while on sabbatical at the University of Minnesota, Department of Applied Statistics. The author wishes to thank Lonnie Zeltzer and Deb Fanurik for discussion, collaboration and data; to Roderick Little and Dick Berk for comments; and to the Minnesota School of Public Health, Biostatistics division, for helpful comments at a seminar.

Abstract

This paper presents an analysis of covariance of data from a study of pediatric pain. However, the usual assumptions of constant variance and normality are not met on the original scale. Simultaneous transformation of the response and baseline covariate is used to accommodate assumption violations. Predictive inference is used to provide interpretable inferences which distinguish between observational characteristics of the subjects and manipulatable treatments. Sensitivity and diagnostic tools greatly strengthen the conclusions that can be drawn from this data set. Without the sensitivity analysis, we are left wondering whether the conclusions rest on a true underlying treatment effect or on cases of questionable quality. In spite of a devil's-advocate effort to produce an alternative model which leads to contradictory conclusions, no such model was found.

Key Words: Box-Cox transformation, Censoring, Conditional Predictive Ordinate, Diagnostics, Influential observations, Outliers.

1 Introduction

Pain is a symptom of many diseases and also a negative outcome of medical procedures and treatments. Children with poor pain tolerance may be expected to have further poor outcomes as a result of intolerance to pain. Avoidance of future medical care is one extreme possible outcome. The long term goal of the research of which this analysis is a part is to develop methods for identifying and treating children with poor pain tolerance.

Children were recruited from an elementary school located on the campus of UCLA (Fanurik, Zeltzer, Roberts and Blount 1993). Children take part in four cold pressor trials; two per session, with sessions separated by two weeks. The first trial of each session is with the dominant, usually right, arm; the second is with the other arm. Each trial consists of immersing the arm in cold water; this is the cold pressor procedure. The arm is removed when the pain is no longer tolerable. Pain tolerance is measured as the duration in seconds of arm immersion. Children were classified into two groups during the first visit by their method of coping with the pain. Coping Style (CS) is *attend* (A) if the child pays attention to the arm immersed in cold water or *distract* (D) if the child thinks about other things, such as a recent pleasant experience. The baseline is the second trial on the first day. The second trial on the second day is the response of interest. Prior to the final trial, one of three randomized treatments is administered: a *null* counseling session (N), counseling to *attend* (A) or counseling to *distract* (D). See Fanurik, Zeltzer, Roberts and Blount (1993) for further description of the trial.

The original analysis (Fanurik, Zeltzer, Roberts and Blount 1993) of this data was a classical analysis of covariance on the untransformed data; no diagnostics or sensitivity analyses were used. The ANOVA table is in table 1. Clearly the baseline is important. The main effect of CS is not important, but CS is predictive of baseline tolerance, with distractors showing longer

Source	DF	F	p-value
Baseline	1	37.2	.0001
CS	1	2.3	.13
TMT	2	5.6	.006
TMT*CS	2	5.4	.007
Error	54	MSE = 1060	
Total	60		

Table 1: ANOVA Table.

tolerances than attenders (Figure 1). Treatment (TMT) and the TMT*CS interaction are highly significant. Table 2 of parameter estimates for this problem is unusually easy to interpret; The treatment effect for distract is large and positive, while the interaction attenders taught to distract is large and negative, and essentially the same size as the main effect for CS=distract. The remainder of the effects are small. Multiple comparisons (table 3) of the group effects shows that the Distracters taught to distract are different from all other groups, and none of the others are different from each other. In particular, distractors taught to distract appear to last an extra fifty to sixty seconds over other groups, regardless of baseline.

This data set has several features which make the analysis non-routine in the area of pediatric pain. Children with long tolerances can reasonably be expected to be more variable trial to trial than those with short tolerances. A plot of the data on the log scale, with the $x = y$ line drawn is given in figure 2. On the log scale, the response looks linear, with constant variance.

A second problem is that after 240 seconds of immersion, the child is instructed to remove the arm from the water, and 240 was recorded as the response. Three observations have a censored baseline (id=21, 56, 60) and two have a censored response (id=17, 56). Three potential models of the censoring suggest three different accommodations to this problem. The first is to model the censored times as unknown values known to be longer than 240 seconds. This is the traditional censoring model. The second is to treat

Parm	Est	T	p-val
Intercept	5.18	0.46	.647
Baseline	0.53	6.10	.0001
CS			
Att	9.94	0.67	.503
Dis	0.00	.	. .
TMT			
None	0.50	0.03	.972
Dis	58.55	4.12	.0001
Att	0.00	.	.
TMT*CS			
None Att	-7.66	-0.37	.713
None Dis	0.00	.	.
Dis Att	-61.54	-3.02	.0038
Dis Dis	0.00	.	.
Att Att	0.00	.	.
Att Dis	0.00	.	.

Table 2: Least squares point estimates.

Tmt CS	LSMEAN	i/j	1	2	3	4	5	6
Non Att	28.1	1	.	.88	.78	.0003	.63	.85
Non Dis	25.8	2	.88	.	.66	.0002	.52	.97
Dis Att	32.2	3	.78	.66	.	.0008	.84	.64
Dis Dis	83.8	4	.0003	.0002	.0008	.	.0014	.0001
Att Att	35.2	5	.63	.52	.83	.0014	.	.50
Att Dis	25.3	6	.85	.97	.64	.0001	.50	.

Table 3: Multiple comparisons for group effects. The LSMEAN column assumes that baseline = 37.4 seconds, the mean baseline tolerance.

those children as different from the remainder, and remove them from the data set. The third model argues that the children's arms have gone numb from the cold, that is, that they have fallen into a competing risk. In this last model, had they correctly responded to the cold they would have indeed had long tolerances, but they would not have been as long as 240 seconds. The three models suggest respectively, impute a longer time for the 240 seconds, delete these cases, or impute shorter times for those observations.

A third problem is that case 56 was removed from the original analysis because he always lasted 240 seconds. This case is included in the analyses shown here a question is whether this case is influential or outlying.

A fourth problem is that pediatric pain is concerned with individuals as well as groups. A finding of a small but significant treatment effect is a statement about a population, but the goal is to treat individuals. Individual variability may easily outweigh minor advantages of a particular treatment. Thus the classical approach presented above does not give a complete inference.

I present a Bayesian analysis of this data that accounts for the problems just presented. Data transformation is used to control the heteroskedasticity. Symmetry suggests that the transformations of the baseline and response measurements be identical leading to a combined Box-Cox (1964)-Box-Tidwell (1962) model. Because of the unknown transformation parameter, the model is non-linear, with the attendant difficulties in parameter interpretation (Box and Cox 1982, Hinkley and Runger 1984). A predictive approach gives interpretable inferences.

Prediction is used to unify the analyses throughout the paper. Predictive distributions and summaries are used for inference. Predictive diagnostics are used to identify influential and outlying observations; a sensitivity analysis is performed to assess the impact of the known (censoring) and suspected (outliers, influential observations) possible model misspecifications.

The paper is structured as follows: section 2 contains the inferential data analysis; section 3 contains the sensitivity analysis. The paper closes with discussion. The appendix contains theoretical development of the statistical tools used.

2 Data Analysis

The 6 groups (2 coping styles by 3 treatments) will be identified by two letter abbreviations, with the first letter D or A for coping style and the second D, A, or N for treatment. The data are assumed to follow the transformation model

$$\begin{aligned} r_i^{(\lambda)} &= x_i\beta_1 + b_i^{(\lambda)}\beta_2 + \varepsilon_i, \\ &\varepsilon_i \sim N(0, \tau^{-1}I). \end{aligned}$$

where x_i is a parameterization of the indicators for the 6 groups; B_i is the baseline tolerance; r_i is the response tolerance; β_1 are the treatment parameters; and β_2 is the parameter for the baseline. The transformation is taken as the standard power family transformation

$$r^{(\lambda)} = \begin{cases} \frac{r^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log r & \lambda = 0 \end{cases}$$

with the transformation taken as the standard power family transformation. The data with case diagnostics are given in tables 4 (CS=1, attend) and 5 (CS=2, distract). The case diagnostics are discussed later. Within treatment and coping style, observations are ordered by L_1 and the numbering begins with case 0, as in Lisp. However, rather than take a point estimate for λ, λ is treated as unknown, so as to avoid possible underestimation of variability due to estimating the transformation.

Computations are based on a simple random sample of size 4000 from an approximation to the posterior. Sampling from $p(\beta, \tau|\lambda, Y)$ is straightforward. Sampling from $p(\lambda|Y)$ is more difficult, however, an approximating

normal with mean .142427 and variance .0142605 provided a very good fit to $p(\lambda|Y)$. The approximation is quite good, the L_1 norm between the approximation and the exact distribution is only .018, which was considered too small to worry about. The density $p(\lambda|Y)$ at $\lambda = 0$ is high, however following Rubin (1984) the transformation to normality, which is unknown and the scale of interest which is the original untransformed scale are kept separate. Details of the model and computations are given in Appendix A.

Figure 3 gives predictive distributions for new observations with a baseline tolerance of 24 seconds, the median of the baseline measurements. Figure 3 is a kernel density estimate from predictive samples of size 4000. It is difficult to distinguish amongst the three attender groups AA, AD, and AN; in the distracters group, DD have substantially longer tolerances than DA which is somewhat longer again than DN. Plots for baselines of 6 and 120 seconds show identical structure except for changes in location.

Table 6 summarizes Figure 3 with predictive means and standard deviations for all six groups and baselines of 6, 24 and 120 seconds. From table 6 and figure 3, we see that for a baseline of 24 seconds, the improvement over baseline for DD is only $32(= 56 - 24)$ seconds, not the erroneous full minute of the treatment effect from the parameter estimates. With the transformation model, the improvement, if any, due to the treatment depends upon the baseline. There is also a regression to the mean effect, where the predictive means for a very low baseline of 6 seconds all show a predicted increase in response, while at 120 seconds, only the DD group has a predicted mean longer than the baseline.

Table 7 gives the probability $P(Z_{new,j} > Z_{new,j'}|Y, B)$ that switching a person from treatment j to treatment j' leads to a longer tolerance. This computation is appropriate for covariates which can be manipulated by the experimenter. Separate probabilities are given for each coping style. Each calculation assumes that the baseline measure, coping style and random pre-

ID	CS	TMT	B	R	L_1	100CPO
0	A	A	35.31	11.71	0.284	0.714
33	A	A	11.92	44.72	0.263	0.325
53	A	A	32.84	25.21	0.076	2.263
35	A	A	23.29	20.67	0.074	2.747
7	A	A	19.03	30.37	0.073	1.959
13	A	A	13.47	15.98	0.073	3.605
54	A	A	30.66	38.47	0.064	1.672
3	A	A	23.41	31.38	0.059	2.041
31	A	A	30.84	37.03	0.059	1.775
10	A	A	26.30	28.64	0.050	2.307
34	A	D	12.99	34.76	0.182	0.807
28	A	D	11.42	27.44	0.138	1.423
60	A	D	16.50	11.12	0.128	3.252
57	A	D	23.18	14.16	0.126	2.694
41	A	D	16.44	12.63	0.106	3.506
47	A	D	13.41	21.19	0.067	2.784
21	A	D	42.22	41.44	0.060	1.599
26	A	D	18.13	19.33	0.059	3.197
24	A	D	18.77	20.34	0.057	3.078
59	A	D	27.61	27.00	0.050	2.422
19	A	N	240.00	116.68	0.12	0.592
2	A	N	10.00	8.27	0.118	4.800
23	A	N	6.24	7.13	0.114	6.300
46	A	N	38.85	48.42	0.113	1.050
45	A	N	33.54	22.65	0.075	2.538
16	A	N	20.03	26.82	0.071	2.197
25	A	N	9.63	15.28	0.071	3.778
6	A	N	11.05	13.86	0.070	4.241
55	A	N	11.19	15.51	0.067	3.834
37	A	N	16.87	18.88	0.059	3.286

Table 4: Data and case diagnostics, part 1: attenders. Id number; CS, coping style, A=attend, D=distract; TMT, treatment, A=attend, D=distract, and N=null; x =baseline tolerance, seconds; y = response tolerance, seconds; L_1; $100 * \text{CPO}_i$.

ID	CS	TMT	x	y	L_1	100CPO
27	D	A	10.51	22.80	0.149	1.645
49	D	A	52.01	20.16	0.145	1.733
43	D	A	12.42	8.06	0.140	4.300
56	D	A	240.00	104.50	0.095	0.685
4	D	A	85.91	60.30	0.073	1.129
50	D	A	14.47	14.53	0.070	4.034
44	D	A	12.58	15.63	0.069	3.772
17	D	A	17.53	21.73	0.068	2.724
36	D	A	49.00	43.00	0.067	1.496
11	D	A	23.93	20.00	0.059	3.101
15	D	D	41.72	240.00	0.668	0.001
58	D	D	36.43	180.19	0.402	0.021
9	D	D	11.75	13.29	0.267	1.091
1	D	D	24.22	20.30	0.241	0.721
22	D	D	44.94	35.97	0.185	0.727
38	D	D	25.13	31.04	0.134	1.358
52	D	D	240.00	240.00	0.130	0.307
48	D	D	42.58	48.94	0.093	1.143
8	D	D	41.20	78.00	0.068	0.935
29	D	D	29.51	63.12	0.063	1.131
42	D	D	29.35	55.27	0.051	1.314
30	D	N	88.89	6.67	0.684	0.016
20	D	N	44.16	65.42	0.265	0.191
14	D	N	45.41	44.31	0.140	0.927
12	D	N	41.20	40.78	0.133	1.041
32	D	N	10.12	7.62	0.098	6.543
51	D	N	24.51	12.19	0.091	4.081
5	D	N	18.95	20.35	0.088	2.549
18	D	N	20.29	11.89	0.083	4.512
40	D	N	16.75	14.66	0.069	3.987
39	D	N	38.89	20.90	0.061	2.961

Table 5: Data and case diagnostics, part 2: distracters. Id number; CS, coping style, A=attend, D=distract; TMT, treatment, A=attend, D=distract, and N=null; B =baseline tolerance, seconds; y = response tolerance, seconds; L_1; $100 * \text{CPO}_i$.

CS/TMT	6	24	120
DD	27.1	56.27	146.2
DA	11.5	25.37	74.0
DN	8.6	19.63	58.9
AD	13.0	28.48	82.9
AA	14.3	31.29	87.1
AN	12.3	27.29	78.7

Table 6: Predictive means for the six groups, for baselines of 6, 24, and 120 seconds. Based on samples of size 4000.

dictive person error is the same under either treatment. The probabilities do not depend upon the baseline. These probabilities are approximately equal to the one sided p-value for the hypothesis test that the difference in treatment effects is equal to zero and may be interpreted using similar intuition. Table 7 shows that the distract treatment for distractors is a substantial effect even for individuals. The other treatments and groups do not have a large effect. The most interesting conclusion is the $P(DD > DA) \approx 1$ and $P(DD > DN) = 1.000$. Mathematical details of the predictive inferences are given in Appendix B.

Table 8 gives the probability that a person from any CS-TMT group combination has greater tolerance than someone else from another CS-TMT group. This computation is appropriate for observational characteristics not under the control of the experimenter. The results do not depend on the particular baseline tolerance. The differences between tables 7 and 8 show how the predictive approach distinguishes between randomized treatments versus observational studies; the inference is much stronger for the treatments than for observational characteristics.

	Coping Style			
	Distracters		Attenders	
Treatment	Attend	None	Attend	None
Distract	0.99925	1.000	.35875	.59075
Attend		0.8355		.7285

Table 7: Posterior probability that placing a person on the row treatment will lead to greater tolerance than putting that same person on the column treatment. Based on samples of 4000.

	dd	da	dn	ad	aa	an
dd	—	0.840	0.900	0.798	0.768	0.817
da		—	0.615	0.438	0.397	0.466
dn			—	0.327	0.291	0.353
ad				—	0.458	0.528
aa					—	0.570

Table 8: Posterior probability that placing a person from the coping style and group from the row has greater tolerance than a different person from the coping style and treatment group given in the column headings. The ordering of the groups is the same as in table 6, dd $\gg$ aa $>$ ad $>$ an $>$ da $\gg$ dn. Based on a sample of 4001.

3 Sensitivity Analysis

This section explores the robustness of the conclusions of the previous section to changes in model specification. Two kinds of model changes are explored, case deletion using case diagnostics, and a sensitivity analysis to determine the effect of altering the likelihood contribution of the censored cases with baseline or response tolerances of 240 seconds.

The purpose of this sensitivity analysis is to search through a large space of a priori plausible alternative models at low cost. If no models are found which are both influential and supported by the data, then we stand by the basic model and its conclusions. Otherwise we identify a small set of extreme models which will be fit using the data. The important conclusions will be recomputed and compared with those from the original model. If the main conclusions are similar to those from the basic model, then we can be much more comfortable in the conclusions. If conclusions are qualitatively different, we conclude that the data does not support strong conclusions. An alternative to the sensitivity analysis is to fit a mixture model encompassing all possible perturbations to the basic model. This requires substantially more work, and will be sensitive to the prior probabilities specified in the mixture model.

First, case diagnostics, the conditional predictive ordinate (CPO) outlier statistic, and L_1 norm case influence statistic, were calculated; these values were given in tables 4 and 5. The CPO statistic was first proposed by Geisser (1980) and has been analyzed by Geisser (1987, 1989, 1990), Pettit (1985, 1990) and Weiss (1993). The CPO_i is the Bayes factor in favor of the original model against the model which deletes case i. The smaller CPO is, the more outlying is the observation. A factor of 10 difference in two values of CPO indicates that one observation is 10 times more likely to be an outlier than another observation. The CPO statistic can be converted to the $(0, 1)$

probability scale, and this quantity is reported in the last columns of tables 4 and 5.

The L_1 influence diagnostic is one of Csiszár's (1967) divergences, as is the Kullback divergence, and is easier to interpret (Weiss 1993). Appendices D and E include a short technical discription of CPO and L_1. Here I briefly describe how to interpret them. The L_1 statistic ranges from 0 to 1. When it is zero, then there is no influence. An L_1 of less than .1 is considered to be relatively uninfluential, around .3 is moderately influential, above .5 is high, and an L_1 of 1 indicates that the posterior from the case deleted model does not share any support with the basic model. The L_1 influence measure is a global measure, and is susceptible to influence on aspects of the posterior that may not be of direct interest.

Three cases $30, 58, 15$ are identified as outlying by CPO, and these three are also very influential by L_1, with values over .4 for each case. For the next stage of the analysis, these three observations were deleted as a set to form one perturbation. Case 52 was neither influential nor outlying by itself.

Next I assess influence of the values for the censored observations. Several children kept their arms immersed for the maximum permitted time of 240 seconds, the baseline is 240 seconds for cases 19, 52, and 56; the response is 240 for cases 15 and 52; additionally, $y_{58} = 180.19$. The next largest observation is $y_{19} = 116.68$. The standard censoring model treats these observations as conditionally normal given the parameters, but with unobserved values known to be larger than 240. However, this model was considered to be potentially inappropriate here, since the reason for the long tolerances is probably that the arms become numb; the children are no longer responding to pain or cold as a stimulus. More appropriate might be to take 240 as an upper bound on the responses. As a lower bound, we might pick either the largest, or second largest uncensored observation, $y_{58} = 180.19$ or $y_{19} = 116.68$. Since y_{19} is only 12 seconds larger than the next observation,

while 180.19 was a minute larger than 116.68, it was also considered as a possible numbness induced measurement. To assess the influence of alternative values, values of 120, 150, 180, 240, and 300 seconds were substituted for the censored values. For values of 120 and 150, case 58 was both perturbed and unperturbed.

Two multiple case deletions were considered, the four cases with baseline or response of 240 and the three influential outliers from the single case diagnostics.

Information about these model perturbations are given in table 9, with L_1 influence statistic and CPO^{-1} outlier statistics. For different values of the perturbed maximal tolerances, CPO can be sensibly be compared to each other. For the case deletion values, the CPO values have been adjusted to make them comparable to each other and to the other values in the table. Appendix C explains about the case deletion adjustments of CPO. The smaller CPO is, the more the data prefer that perturbation to the original unperturbed model.

The perturbation of values from 240 to 300 is actually less supported by the data than the null model since CPO > 1, but the other perturbations from 240 to shorter values are more supported than the null. The value of CPO decreases smoothly as the perturbed tolerance measure decreases towards 120, and as all 5 observations with baseline or response greater than 120 are changed to 120. All of the changes are influential, and the influence is increasing with decreasing CPO.

Two further analyses are suggested by the sensitivity analysis so far: one deleted the three outlying cases, and one where the large x and y values are perturbed to 120 seconds. These two perturbations are quite different. In the families of perturbations considered, these are the most influential and the most outlying. If the major conclusions do not change under these perturbed models, then we can be more comfortable that the major conclusions are not

Type of perturbation	No. of cases	cases	new value	L_1	CPO
Increase 240's	4	15, 19, 52, 56	300	.3	8.12
No perturbation	0			0	1
decrease 240's	4	15, 19, 52, 56	180	.365	.077
decrease 240's	4	15, 19, 52, 56	150	.511	.022
decrease 240's	4	15, 19, 52, 56	120	.597	.0090
decrease 240's and 180	5	15, 19, 52, 56, 58	150	.648	.0044
decrease 240's and 180	5	15, 19, 52, 56, 58	120	.805	.00028
delete cases	3	15 30 58		.944	3.94e-8
delete cases	4	15, 19, 52, 56		.664	2.32e-4

Table 9: Summary of perturbation results. Lists type of perturbation; number of cases involved; case numbers; new baseline or response value for changed 240 and 180 values; L_1 influence statistic; CPO outlier statistic. For the case deletion perturbations, the CPO statistic has been adjusted as discussed in Appendix C.

sensitive to the assumptions either about the cases with $y = 240$ or $x_{2i} = 240$ or with the outliers left in the model. Interest lies especially in checking table 4 for these perturbed models. The results are given in table 7. There are not major changes in the outcomes of comparing distracters to the other two groups. The largest change seems to be for comparing da to dc which is not of great interest because both treatments are second rate treatments.

4 Discussion.

The predictive inference tools provide interpretable inferences which distinguish between observational characteristics of the subjects and manipulatable treatments.

The sensitivity and diagnostic tools greatly strengthen the certainty of the conclusions that can be drawn from this data set. In spite of the devil's-advocate efforts to produce an alternative model which leads to contradictory conclusions, no such model was found. Without the diagnostic and sensitiv-

		Model		
				240 →
CS	TMT > TMT	Orig	Del3	120
DIS	DIS > ATT	1.0	1.0	.99
	DIS > NON	1.0	1.0	1.0
	ATT > NON	.84	.56	.91
ATT	DIS > ATT	.36	.36	.37
	DIS > NON	.59	.65	.50
	ATT > NON	.73	.77	.63

Table 10: Predictive probabilities that one treatment is better than another under the original and two perturbed models, the delete 3 outliers model, and the model that adjusts all large observations down to 120 seconds. Based on samples of size 2000.

ity analysis, we are left wondering whether the conclusions rest on a true underlying treatment effect or on the cases of questionable quality.

One shortcoming of the sensitivity analysis presented here is that the different plausible perturbations define separate models, and the possibility of joint perturbation has not been considered. This in principle invites a combinatorial explosion of models as models with 2, 3 and more outliers are considered, plus combinations with other perturbations. An infinite regress of analyses is always possible and the analysis shown here was deemed satisfactory for this data set since the sensitivity analyses strongly support the conclusions from the basic analysis.

An additional value to the sensitivity is for design of future trials. Future studies may have the children remove their arms after three or even two minutes. The sensitivity analysis shows that while details of the conclusions will change, the overall qualitative conclusions will not.

References

[1] Box, G. E. P. (1980). Reply to comments on "Sampling and Bayes'

inference in scientific modelling and robustness" *JRSS-A*, 143, 425- 430.

[2] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion). *JRSS-B*, 26, 211-252.

[3] Box, G.E.P. & Cox, D.R. (1982). An analysis of transformations revisited, rebutted. *JASA*, 77, 209-210.

[4] Box, G.E.P. & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4, 531-550.

[5] Carlin, B. P. & Polson, N. G. (1991). An expected utility approach to influence diagnostics. *JASA*, 86, 1013-1021.

[6] Carlin, B. P. & Polson, N. G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4*, pp. 577-586, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, (Eds). Oxford: Clarendon Press.

[7] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2**, 299-318.

[8] Fanurik, D., Zeltzer, L. K., Roberts, M. C. and Blount, R. L. (1993). The relationship between children's coping styles and psychological interventions for cold pressor pain. *Pain*, in press.

[9] Geisser, S. (1980). Comments on "Sampling and Bayes' inference in scientific modelling and robustness" *JRSS-A*, 143, 416- 417.

[10] Geisser, S. (1987). Influential observations, diagnostics and discordancy tests. *J. Applied Statistics*, 14, 133-142.

[11] Geisser, S. (1989). Predictive discordancy tests for exponential observations. *Canadian Journal of Statistics*, 17, 19-26.

[12] Geisser, S. (1990). Predictive approaches to discordancy testing. In *Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard*, S. Geisser;J. S. Hodges, S. J. Press, A. Zellner, (Eds), pp. 321-335. Amsterdam: Elsevier-North Holland.

[13] Geisser, S. (1991). Diagnostics, divergences and perturbation analysis. In *Directions in Robust Statistics and Diagnostics I*, W. Stahel and S. Weisberg, (Eds), pp. 89-100. Springer-Verlag, New York.

[14] Geisser, S. (1992). Bayesian perturbation diagnostics and robustness. In *Bayesian analysis in statistics and econometrics, Lecture notes in Statistics, Vol. 75*, P. K. Goel and N. S. Iyengar, (Eds)., pp. 298-301.

[15] Gelfand. A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398-409.

[16] Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data (with discussion). *JASA*, 79, 302-320.

[17] Kass, R. E., Tierney, L. & Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, **76**, 663-674.

[18] Pettit, L. I. & Smith, A. F. M. (1985). Outliers and Influential Observations in Linear Models. In *Bayesian Statistics 2*, Eds. J. M. Bernardo, M. DeGroot, D. Lindley, and A. F. M. Smith, pp. 473-94 Amsterdam: North Holland.

[19] Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *JRSS-B*, 52, 175-184.

[20] Rubin, D.B. (1984). Comments on "The analysis of transformed data". *JASA*, 79, 309-312.

[21] Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician*, 41, 169-178.

[22] Stigler, S. M. (1980). Comments on "Sampling and Bayes' inference in scientific modelling and robustness". *JRSS-A*, 143, 424- 425.

[23] Weiss, R.E. (1992). Bayesian sensitivity analysis using divergence measures. Submitted for publication.

[24] Weiss, R. E. (1993). Identifying outlying perturbations. University of Minnesota, School of Statistics Technical report # 594.

A The Model

The general model considered here is

$$\begin{aligned} Y^{(\lambda)} &= X_1\beta_1 + X_2^{(\lambda)}\beta_2 + \varepsilon, \\ \varepsilon &\sim N_n(0, \tau^{-1}I) \end{aligned} \tag{1}$$

where $Y^{(\lambda)}$ is a vector of length n with i^{th} element $y_i^{(\lambda)}$, a transformation of y_i parameterized by λ. Similarly, $X_2^{(\lambda)}$ is a baseline measure with individual elements $x_{i2}^{(\lambda)}$; $X_1, n \times p$, a matrix of fixed covariates; β_1 is a vector of regression coefficients; β_2 is a scalar regression coefficient; and ε is a vector of iid normal errors. The analysis of data with power transformations of both response and predictor variables is not new, especially in the Econometrics literature (see Sakia 1992 for a review). This section treats the analysis of (1) generally, while the next section analyzes the pain data. For the pain data, X_1 is a parameterization of the design matrix of a 2 by 3 analysis of variance, while X_2 is the baseline measure. Individual baseline measures will be denoted by x_{2i}.

Let $\beta^T = (\beta_1^T, \beta_2), X = X(\lambda) = (X_1|X_2^{(\lambda)})$. The dependence of X on λ will often be suppressed for ease of notation. The likelihood resulting from (1) is

$$L(\beta, \tau, \lambda) \propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\left(Y^{(\lambda)} - X(\lambda)\beta\right)^t \left(Y^{(\lambda)} - X(\lambda)\beta\right)\right\} J(Y, \lambda)$$

where

$$J(Y, \lambda) = \prod_{i=1}^{n} J(y_i, \lambda) = \prod \partial y_i^{(\lambda)} / \partial y_i$$

is the Jacobian of the transformation. The maximizers of $L(\beta, \tau, \lambda)$ for fixed λ are the usual maximum likelihood estimates

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T Y^{(\lambda)}$$

and

$$\hat{\tau}(\lambda) = n(\mathrm{RSS}(\lambda))^{-1}$$

where $\mathrm{RSS}(\lambda) = (Y^{(\lambda)} - X\hat{\beta}(\lambda))^T(Y^{(\lambda)} - X\hat{\beta}(\lambda))$. The profile log likelihood of λ becomes

$$\ell(\lambda) = \ell(\hat{\beta}(\lambda), \hat{\tau}(\lambda), \lambda) = \frac{n}{2}\log\left(\frac{n}{\mathrm{RSS}(\lambda)}\right) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\log y_i \,.$$

This can be maximized and plotted as in Box-Cox (1964), and asymptotic χ^2 test based on $2(\ell(\hat{\lambda}) - \ell(\lambda_0))$ can be used to test $H_0 : \lambda = \lambda_0$.

Bayesian inference in this model permits simple graphical and numerical posterior summaries, including summaries that can replace frequentist multiple comparison procedures. The posterior

$$p(\lambda, \beta, \tau|Y) = p(\lambda|Y)p(\tau|\lambda, Y)p(\beta|\tau, \lambda, Y)$$

can be explicitly calculated up to the normalizing constant for $p(\lambda|Y)$ which, since λ is often a scalar, can be dealt with quite easily. Here we take a uniform improper prior $p(\beta, \tau, \lambda) \propto \tau^{-1}$. The regression coefficients β are conditionally normal:

$$[\beta|\tau, \lambda] \sim N\left(\hat{\beta}(\lambda), (\tau X^T X)^{-1}\right) \,, \tag{2}$$

while τ given λ has a gamma distribution

$$[\tau|\lambda, Y] \sim \Gamma\left(\frac{n-p}{2}, \frac{\mathrm{RSS}(\lambda)}{2}\right) \tag{3}$$

with density $p(\tau|\lambda, Y) \propto \tau^{\frac{n-p}{2}-1}\exp\left\{-\frac{\mathrm{RSS}(\lambda)\tau}{2}\right\}$ and

$$p(\lambda|Y) \propto |X^T X|^{-1/2}(\mathrm{RSS})^{-(\frac{n-p}{2})}(\prod y_i)^{\lambda-1} \,. \tag{4}$$

Sampling based Bayesian inference (Gelfand and Smith 1990) is used for the computations in this model. First sample from $p(\lambda|Y)$, then from $p(\tau|\lambda, Y)$ and finally from $p(\beta|\lambda, \tau, Y)$. Sampling from $p(\lambda|Y)$ can be accomplished either by importance sampling or discrete approximation. In the pain data a normal approximation to $p(\lambda|Y)$ suffices. Repeating the sampling M times produces M independent samples $\theta^{(m)} = (\beta^{(m)}, \tau^{(m)}, \lambda^{(m)}), m = 1, \ldots M$ from the posterior.

B Predictive Inference.

The predictive distribution of Z at fixed x_f is

$$f(z|Y,x_f) = \int\int\int \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left\{-\frac{\tau}{2}\,(z^{(\lambda)} - x_f^T\beta)^2\right\} J(z,\lambda)p(\beta,\tau,\lambda|Y)d\beta d\tau d\lambda\ .$$

The dependence of x_f^T on λ is suppressed. Predictive distributions are simulated by sampling w from a $N(x_f^T\beta^{(m)}, (\tau^{(m)})^{-1})$, then backtransforming w into $z_f^{(m)}$. When the transformation is the standard power family transformation

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y_i & \lambda = 0 \end{cases}$$

then

$$z_f^{(m)} = \begin{cases} (\lambda w + 1)^{1/\lambda} & \lambda \neq 0 \\ e^w & \lambda = 0 \end{cases}$$

If $(\lambda w + 1) \leq 0$ then resample w again. Like the Box-Cox model, it is assumed that the response is non-negative and the censoring has negligible impact on the inference.

Inference for comparing different treatments or treatment combinations is done in a predictive graphical fashion, and with appropriate summary statistics. This has several advantages over usual point estimation and testing procedures. It avoids the uninterpretability of marginal estimated coefficients and standard errors (Box and Cox 1982, Hinkley and Runger 1984). It also permits assessment of practical differences on the original scale of measurement, yet permits uncertainty due to the transformation to be propagated. The predictive distribution on the original measurement scale can be a great aid in communicating with subject matter researchers who have difficulty interpreting log, square root and other transformed scales. In the pain data, the graphical predictions clarify the ambiguous information in the ANOVA table, which featured one interesting but weak main effect and an interesting but weak interaction. Finally, using a predictive framework makes all

parameterizations of the ANOVA model equivalent, and we need not worry about proper interpretation of coefficients. Relevant summary measures can easily be estimated, for example means and standard deviations of predictions and for differences in predictions. Taking a predictive approach permits assessment of practical as well as statistical impact of treatments.

The predictive distribution of $z|x_f$ for observations in different treatment groups may be computed for equal values of the baseline covariates to assess treatment differences. We also may be interested in the difference in response of either the same or two different people in two different groups denoted by x_{f1} and x_{f2}. Let z_{f1} and z_{f2} be conditionally independent predictions at x_{f1} and x_{f2}. A plot of the densities of z_{f1} and z_{f2} is the current best guess as to the responses of two people with covariates x_{f1} and x_{f2}. One can also consider the posterior of $z_{f1} - z_{f2}$. These densities are relevant when the two groups are distinguished by unchangeable characteristics of the individuals. A summary measure of the difference between z_{f1} and z_{f2} of particular interest is

$$\begin{aligned} P(z_{f1} > z_{f2}) &= P((x_{f1}^t\beta + \epsilon_{f1})^{(\lambda)} > (x_{f2}^t\beta + \epsilon_{f2})^{(\lambda)} \\ &= P((x_{f1} - x_{f2})^t\beta > \epsilon_{f1} - \epsilon_{f2}). \end{aligned}$$

Suppose $(x_{f1}-x_{f2})^t\beta = \beta_j$, as happens in the pain data when we are comparing two predictive distributions that differ only by a single indicator variable; the baseline measurements are assumed equal. Then

$$P(z_{f1} > z_{f2}) = E\left[\Phi\left(\frac{\beta_j}{2^{.5}\sigma}\right)\right], \tag{5}$$

where $\Phi(a)$ is the standard normal cumulative distribution function. The probability (5) can be approximated by

$$P(z_{f1} > z_{f2}) \approx \Phi\left(\frac{\hat{\beta}_j}{(\mathrm{Var}(\hat{\beta}_j) + 2\hat{\sigma}^2)^{.5}}\right).$$

When the difference between x_{f1} and x_{f2} is a treatment which can be modified by the experimenter, we may be interested in a different comparison. Let z^*_{f1} and z^*_{f2} be corresponding predictions at x_{f1} and x_{f2} whose unobserved normal errors ϵ^*_{f1} and ϵ^*_{f2} are equal with probability 1. Then $P(z^*_{f1} > z_{f2}|Y)$ is the probability that treatment 1 produces a larger response than treatment 2 on one person and, again assuming $(x_{f1} - x_{f2})^t\beta = \beta_j$

$$P(z^*_{f1} > z^*_{f2}|Y) = P(\beta_j > 0|Y) \approx \Phi\left(\frac{\hat{\beta}}{(\text{var}(\hat{\beta}_j))^{.5}}\right) \tag{6}$$

for monotone transformations and error structures that are location-scale models after transformation, providing a predictive interpretation to the coefficient β_j. The approximation is equal to the classical one sided p-value for testing $H_0 : \beta_j = 0$. The probabilities (5) and (6) do not depend on the baseline measurements. From (6) compared to the approximation to (5) we see that we have a stronger inference for treatment effects than for the coefficients of observed variables.

Treatment and group difference assessments can be evaluated predictively by taking predictive means within groups as well as on the individual level given here. For example one could compute the probability that the average response on treatment 1 is greater than the average response on treatment 2. If substantive differences are found for individuals, then necessarily there will be differences for groups; thus an analysis that shows effects for individuals will show effects for groups. The converse need not hold. Group level evaluations may be pertinent for global decisions by policy makers, however individual differences in utility can easily outweigh global policy recommendation, something less likely to happen when a treatment has a beneficial effect that is noticeable on the individual level.

C Outlier Diagnostics.

Diagnostics can and should be implemented in any new model. The tools used here are extensible to other models. It is of specific interest to identify outliers in pediatric pain research: these children may have low or high tolerance and may be at risk for (not necessarily respectively) low or high usage or avoidance of medical procedures. However, the outliers of interest are from marginal models where the distribution of the baseline measure is also modeled; not outliers conditional on the baseline measure as in model (1); if both the baseline and response are small, the case will not be identified as an outlier yet these children are specifically of interest for further intervention. In the current analysis, outlier statistics are used to identify observations which may be discordant and may affect the analysis.

A Bayesian outlier statistic is CPO, the conditional predictive ordinate of Geisser (1980, 1987, 1989, 1990), Pettit and Smith (1985), Pettit (1990) and Weiss (1993). Define

$$\text{CPO}_i = f(y_i|Y_{(i)}) = \int f(y_i|\theta, x_i)p(\theta|Y_{(i)})d\theta,$$

which is the normalizing constant in the updating version of Bayes theorem

$$p(\theta|Y) = \frac{p(\theta|Y_{(i)})f(y_i|\theta)}{\text{CPO}_i} \tag{7}$$

and also a computational byproduct of the influence diagnostics of the next section. Carlin and Polson (1991, 1992) estimate CPO_i directly from (7), requiring n samples from each of the n densities $p(\theta|Y_{(i)})$. By solving (7) for $p(\theta|Y_{(i)})$, CPO can be computed as

$$\text{CPO}_i^{-1} \approx M^{-1} \sum_{m=1}^{M} [f(y_i|\theta, x_i)]^{-1}.$$

which only requires a single posterior sample from $p(\theta|Y)$.

The diagnostic CPO is dependent upon the scale of analysis (Box 1980, Geisser 1980, Stigler 1980); this is solved by choosing a particular measurement scale for analysis: for the pain data, the original scale in seconds is chosen. Changing the analysis scale from seconds to hours also causes a change in the absolute level of CPO_i, this can be normed internally by realizing that most of the data is good data, and that most of the CPO statistics are representing good data, not bad. The lower quartile of CPO is .01, which is a convenient number to use: all CPO statistics for case deletion have been multiplied by 100 in tables 4 and 5. In data sets with non independent and identically distributed data, the maximum possible value $\sup_z f(z|x_i, Y_{(i)})$ for any one observation i is dependent on the covariates. In normal theory multiple linear regression with known σ, this maximum value is $((1-h_i)/(2\pi h_i \sigma^2)^{-1})^{.5}$. For this data set and after transforming the baseline variables by the mean value of λ, the maximum value varied only by a factor of 2, which was relatively unimportant given the wide range of the CPO_i themselves.

For multiple case deletion, with only a few specially select sets of deleted cases, it is less easy to provide an internal norming of the values. A small simulation was undertaken, where 100 sets of 4 observations were deleted, and CPO computed. The five number summary (min, lower quartile, median, upper quartile and max) of CPO from this sample was (2.43e-11, 2.21e-08, 1.01e-07, 3.55e-07, 2.97e-06). The set of four observations with censored x and y values have a CPO that is 10.5 times smaller than the smallest CPO in the 100 sets of 4, suggesting that this set is outlying. To assess the outlyingness of the three case deletion, first note that the 25^{th} percentile of the single observation CPOs is .01, while the same percentile of the sets of 4 is 2e-8, or approximately the fourth power of the single observation CPOs, suggesting that for sets of 3 observations, approximately, a typical value of CPO could be expected around 1e-6. In contrast, $CPO_{15,30,58} \approx$ 4e-14, a

factor of 4e-8 smaller than 1e-6 and 2000 times smaller than the smallest CPO from the sets of four. The values of CPO have already been adjusted in table 9 for the two case deletion perturbations.

D Influence Analysis.

Influential cases can be identified using the L_1 distance influence statistic of Weiss (1992). See also Geisser (1991,1992).

$$L_{1i} = \frac{1}{2}\int |P(\theta|Y) - P(\theta|Y_{(i)})| d\theta$$

where the parameter vector $\theta = (\beta, \tau, \lambda)$ and $p(\theta|Y_{(i)})$ is the posterior density given the data omitting the i^{th} case. The statistic L_{1i} is bounded between 0 and 1; 0 indicating no influence and 1 indicating that the two posteriors $p(\theta|Y)$ and $p(\theta|Y_{(i)})$ do not share support. If interest lies in a particular parameter or predictor L_{1i} can be replaced by

$$L_{1i,\gamma} = \frac{1}{2}\int \left|p(\gamma(\theta)|Y) - p(\gamma(\theta)|Y_{(i)})\right| d\gamma \ .$$

However $0 \leq L_{1i,\gamma} \leq L_{1i} \leq 1$, so that L_{1i} small guarantees $L_{1i,\gamma}$ small. Also if a vector of predictions $\underline{z}_m$ are of interest then taking $\theta^* = (\beta, \tau, \lambda, \underline{z}_m)$ and

$$L^*_{1i} = \frac{1}{2}\int \left|p(\theta^*|Y) - p(\theta^*|Y_{(i)})\right| d\theta^*$$

then

$$L^*_{1i} = L_{1i}$$

and so $L_{1i} \geq L_{1i,\underline{z}_m} \geq 0$. In fact, an even stronger statement is possible which gives L_1 a predictive interpretation. Let $\underline{z}_m$ be an set of m predictions for each m and assume that the ratio $f(y_i|\underline{z}_m, Y_{(i)})[f(y_i|\theta)]^{-1}$ converges pointwise almost everywhere to 1. Then

$$L_{1i,\underline{z}_m} \nearrow L_{1i}.$$

In other words, if the limiting information in the predictive densities is equivalent to knowing the parameter θ, then the limiting influence on the predictions is equal to the influence on the posterior.

If interest lies in a posterior or predictive expectation $E[\gamma(\theta^*)|Y]$, then influence may be assessed by the quantity

$$\begin{aligned} d_i &= E\left[\gamma(\theta^*)|Y_{(i)}\right] - E\left[\gamma(\theta^*)|Y\right] \\ &= \int \gamma(\theta^*)\left[P(\theta^*|Y_{(i)}) - p(\theta^*|Y)\right] d\theta^* \\ &= E\left[\gamma(\theta^*)(\frac{\text{CPO}_i}{f(y_i|\theta, x_i)} - 1)\right] . \end{aligned}$$

so

$$d_i = \text{CPO}_i * \text{Cov}\left(\gamma(\theta^*), [f(y_i|\theta, x_i)]^{-1}\right) . \tag{8}$$

If γ is an indicator function, then $d_i = L_{1i,\gamma}$. Table 10 permits us to compute d_i for γ's of particular interest.

E General Perturbations

Several observations have baseline x_i or response y_i values censored at 240 seconds. The appropriate values are difficult to model, instead a sensitivity analysis will be undertaken to see if deleting these observations as a set or changing their values leads to a change in one of the major conclusions of the analysis. Define a perturbation function (Weiss 1992; Kass, Tierney and Kadane 1989) for example by

$$h(\theta) = \frac{f(y_i + \omega|x_i, \theta)}{f(y_i|x_i, \theta)} ,$$

to assess the influence of changing y_i to $y_i + \omega$. Then the perturbed posterior is $p_h(\theta|Y) = p(\theta|Y)h(\theta)[E[h(\theta)]^{-1}$. The influence of this perturbation can be assessed by L_{1i} or d_i, and the previous formulas and interpretations apply directly without modification except that the perturbation has changed from

single case deletion to y_i or x_i perturbation. The perturbations can also be more complicated than $h(\theta)$ above; most perturbations in table 9 are combinations of x and y perturbations for four or five cases.

The CPO statistic can also be computed for these perturbations (Weiss 1993). Consider the perturbed posterior as resulting from a perturbed model, then CPO is the Bayes factor against the perturbed model in favor of the original model.

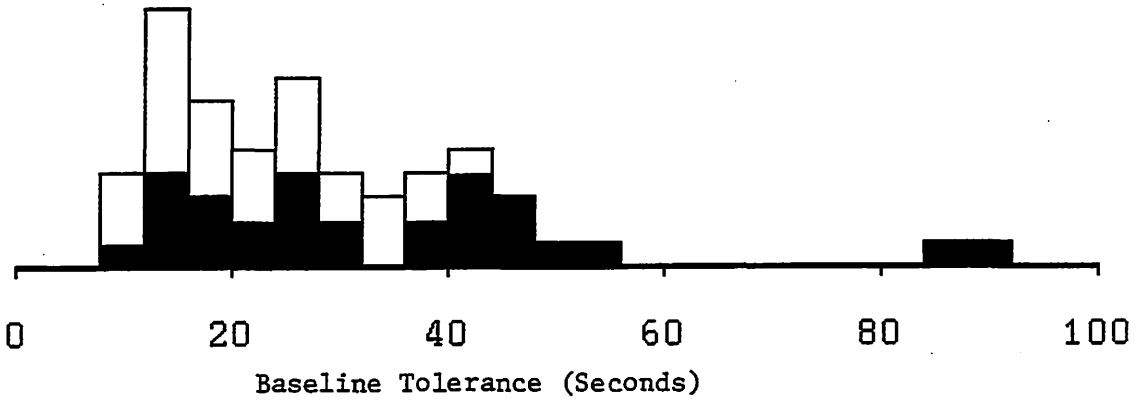

Figure 1. Histogram of baseline tolerance in seconds. Each histogram bar is comprised of solid or clear blocks, one for each case. Solid blocks correspond to distractors, clear blocks to attenders. Three baseline measures of 240 seconds (1 attender, 2 distractors) have been removed to permit more detail.

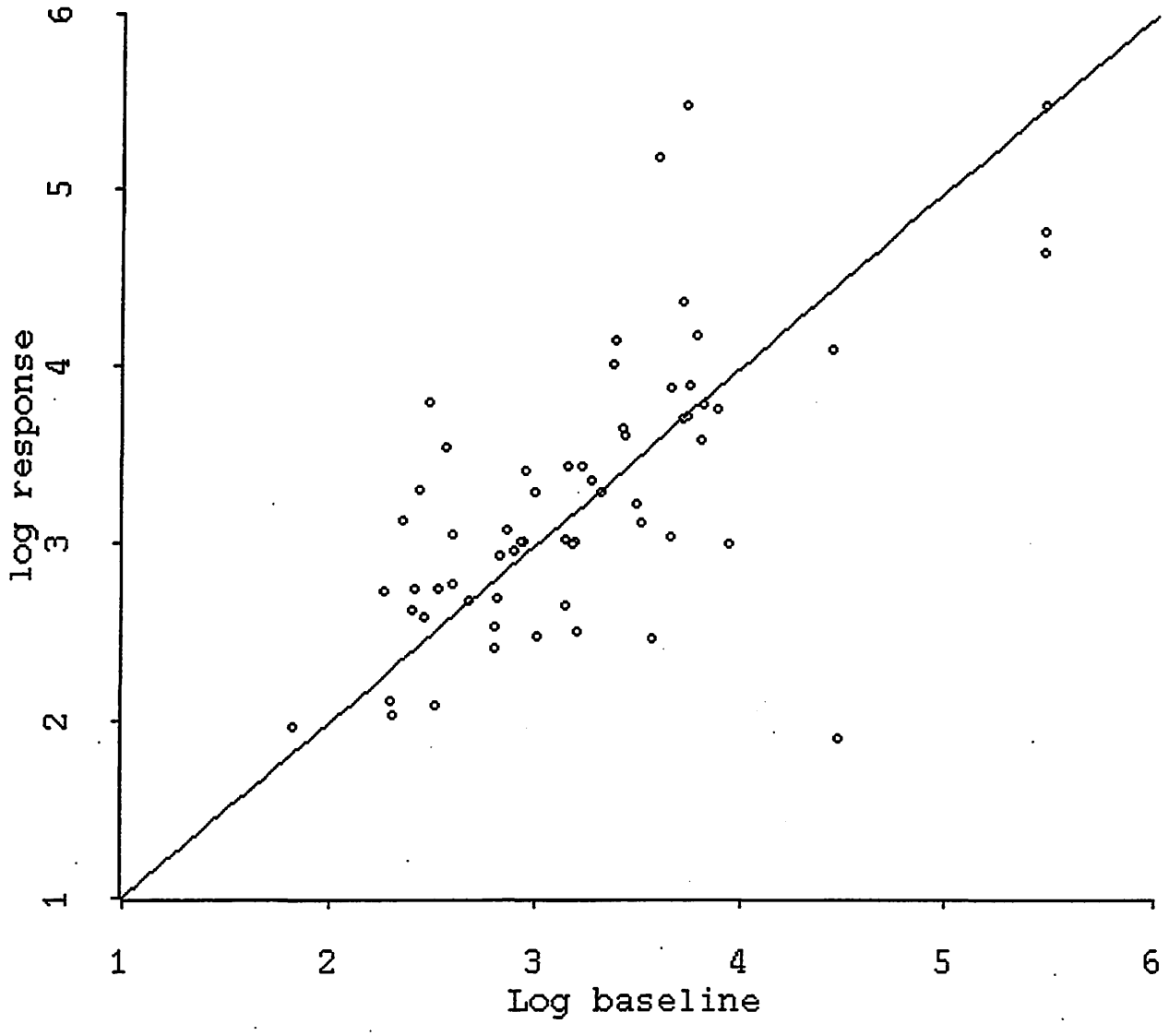

Figure 2. Plot of log response Y vs. log baseline X. The $x = y$ line is plotted. The response seems to have constant variance.

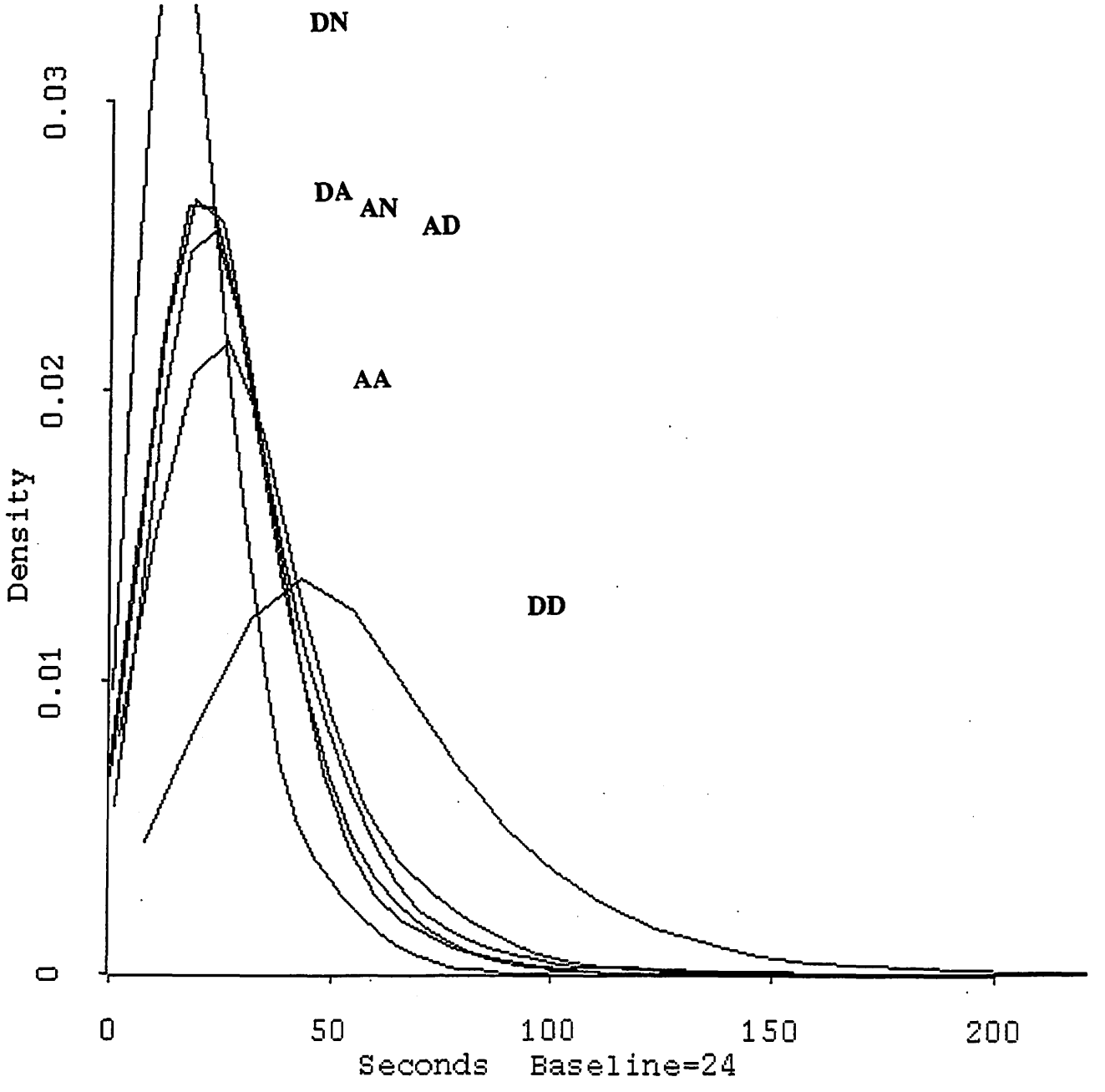

Figure 3. Kernel smoothed plots of predictive distributions of new observations for each of the 6 groups, given a baseline measurement of 24 seconds.